# Geotagging Flickr Photos And Videos Using Language Models

Sanket Kumar Singh
University of Alberta
Edmonton, AB, Canada
sanketku@ualberta.ca

Davood Rafiei
University of Alberta
Edmonton, AB, Canada
drafiei@ualberta.ca

## ABSTRACT

This paper presents an experimental framework for the Placing tasks, both estimation and verification at MediaEval Benchmarking 2016. The proposed framework provides results for four runs - first, using metadata (such as user tags and title of images and videos), second, using visual features extracted from the images (such as tamura), third, by using the textual and visual features together and fourth, using metadata as in the first run but with the training data augmented with external sources. Our work mainly focusses on textual features where we develop a language-based model using bag-of-tags with neighbour based smoothing. The effectiveness of the framework is evaluated through experiments in the placing task.

## 1. INTRODUCTION

The goal of this work is *to estimate* the coordinates of an image or a video on the world map and *to verify* whether an image belongs to a given location. Tags assigned to a photo may not be location-specific and even the location-specific tags can be vague and may refer to multiple locations. Some photos have no tags or have only tags that have not been seen before (e.g. in the training phase). All these issues make location prediction from user tags challenging. We address these problems by learning the associations between user tags and locations and by using this information in our prediction.

## 2. RELATED WORK

Language modeling is used in placing photos on a map. In particular, Pavel et. al [7] place a grid of fixed degree over the world map and map train instances to cells based on their coordinates.They learn a model which allows them to predict the location of the test instances on the grid. Though this work provides several smoothing techniques to predict the location of a test instance whose tags are not seen, it does not differentiate between general and location specific tags. Giorgos et. al in [4] use a similar model but capture information regarding how many users use a particular tag in a particular region. Additionally, they use Shannon's Entropy to give small weights to tags which are user specific or general. Our base model is the same, as it provides a weighting of each tag based on its popularity among users in describing a place, but we experiment with additional components to boost the performance. Another related weighting scheme is that of Aibek et. al in [6] which uses the Kullback-Leibler divergence to differentiate between class-specific and general terms. Even though that work is done in a different context, we experiment with this model in identifying location specific tags.

## 3. PROPOSED APPROACH

The proposed framework consists of two phases: (1) preprocessing the placing dataset [1], and (2) building the model and doing the predictions.

**Preprocessing** Each photo or video has a title, some user tags and the id of a user who posted it. After removing punctuations and special characters from the title, the remaining terms are included in the tag set. This helps the cases where a photo has a title but no tags. In Run 4, which is also based on textual metadata, we include in our training photos instances extracted from the YFCC100M [9] dataset which are uploaded by users other than those in our test set. Furthermore, we augment the tag set with place names from Geonames [10] and assign the location tags to cells based on location coordinates. In all run, each tags that is used by only one user is removed to reduce noise, and the remaining tags are then used for training. For testing, we only use user tags in each run except for Run 4, where we additionally use title and description, for those test instance which have no user tag or none of the tag are found in train data. Our goal in Run 4 is to use as much data as possible. To build a model for Run 2 (which uses visual features), we use 2,182,400 images with Tamura [8] features; the features are preprocessed so they can be fed into Vowpal Wabbit (VW) [5], which is used to train the model. The dataset has 2,735 counties and these are used as labels for training; for our training, county was the smallest region with enough data points per label (812 on average compared to 38 for town).

**Methodology** For the estimation task, we place a grid of 1, 0.1 and 0.01 degrees and predict a cell $c$ for each test photo based on a generative model which estimates the probability $p(t_i|c)$ that the tags $t_i$ in the photo are emitted from cell $c$. The model captures the degree at which a tag is popular among users in describing a location within a cell, i.e.

$$p(t_i|c) = \frac{\text{number of user who use tag } t_i \text{ in cell c}}{\text{number of user who use tag } t_i \text{ globally}},$$

$$p(T|c) = \sum_{i=1}^{n} p(t_i|c)$$


*MediaEval 2016 Workshop,* Oct. 20-21, 2016, Hilversum, Netherlands.

where n is the number of tags in a test instance T. A cell $c$ that gives the maximum p(T|c) is considered as predicted cell of the test instance T. We further extend the base model by performing a neighbour based smoothing as in [7], taking into account who use tag t in the neighbouring cells of cell $c$. Since we need to estimate the actual coordinates of a test instance within a cell, we use the coordinate of a training instance in the same cell that has the maximum Jaccard similarity to the test instance.

Test instances that have no tags (or their tags are not seen in the training set) are assigned to the cell with the largest number of training instances. In this case, the coordinates of the training instance which has the minimum Karney's distance [3] to other instances in the cell is considered as the estimated coordinate. To use visual features, we train a one-against-all multiclass model using VW to predict a county for the test instance. The coordinates are estimated using the same strategy as before, based on the coordinates of a training instance. Since textual features provide a more accurate estimation, visual features are used in Run 3 only if a photo has no textual features. Otherwise, only textual features are used. For the verification task, we use the place information of the training instance, used to predict the coordinates in the estimation task, and mark a test instance verified if its predicted location string contains the given place name.

## 4. RESULTS AND ANALYSIS

We performed our experiments for the estimation task using grids of 0.01, 0.1, 1.0 degrees and evaluated the results using precision at each distance, average distance error (ADE), median distance error (MDE) and the verification accuracy (VA). The results are listed in Tables 1 and 2. From Table 1, we can see that the precision for large distances is high as each target cell covers more area and has more tags. Additionally, as we apply our neighbour based smoothing using adjacent cells, more tags from neighbours are included, which is useful in cases where tags cover wider area such as tags with province name or geographical division which cover more than one grid. This results in an improved cell prediction accuracy.

Analyzing the wrong predictions using the validation set, we find that misspellings, mismatches between plural and singular forms, and the differences in spelling (such as "barcellona" for "Barcelona", "nederland" for "Netherlands") are some of the causes for the tags not to be found in a correct cell. Famous spots such as "the Empire State" building in New York are easily predicted because of abundant location specific tags. However, instances with general words such as "bogus" and "finding" lead to prediction of wrong cells. In an experiment comparing top-k and top-1 predictions for test instances, we found that top-10 accuracy was 47.74% while top-1 accuracy was 31.80% (for photos and video together) using 0.1 degree grid (∼10 km). Furthermore, the predicted cells were closer to the real cells. Another set of instances that were difficult to predict were 335845 test instances (including photos and videos) which either had no tags or their tags were not used by any user in the training set. We assign these instances to the most popular cell, which only gives a correct prediction for 3751 instances.

For Run2, we use Tamura features to train a multiclass model using VW. As the dataset consists of different landscapes, animals, places etc., it is difficult to distinguish between different places from where a photo or video is taken and thus model mainly predicts by most popular county. For Run 4, we augment the cells with place names from Geonames (giving it an arbitrary user id) and from YFCC100M dataset. Since the tags which are used by only one user are removed, only Geonames tags which are used by an actual user in the cell are retained. This increases the count of place specific tags which are used by real users. Using title and description for test instances, which have no user tags or their tags are not found in the training set, reduces the median distance error for the estimation task.

Before reaching the proposed approach, we tried to find location specific tags by assessing their frequency concentration in a region, as compared to the whole map. This approach, however, did not work for instances where the same tag was equally present at two or more places, that were far from each other. Further, we used the KL-Divergence to separate probability distribution of general tags from location specific tags but this approach also did not work well as the model ended up giving more weights to user specific tags such as "lehmans", "gladston", etc.

| Run | Media | 10(m) | 100(m) | 1(km) | 10(km) | 100(km) | 1000(km) |
|---|---|---|---|---|---|---|---|
| 1 | photo | 0.27 | 2.88 | 14.13 | 35.28 | 50.28 | 64.17 |
| | video | 0.27 | 3.03 | 13.50 | 33.24 | 47.60 | 60.08 |
| 2 | photo | 0.0 | 2.0 | 0.42 | 2.13 | 4.0 | 22.97 |
| | video | 0.0 | 0.0 | 0.14 | 0.81 | 1.77 | 6.95 |
| 3 | photo | 0.27 | 2.89 | 14.13 | 35.26 | 50.25 | 64.03 |
| | video | 0.27 | 3.03 | 13.50 | 33.24 | 47.60 | 60.08 |
| 4 | photo | 0.27 | 2.94 | 13.24 | 33.02 | 51.14 | 64.58 |
| | video | 0.27 | 3.36 | 13.29 | 32.61 | 49.35 | 61.18 |

Table 1: Precision at different distances (in %)

| Run | Media | ADE(km) | MDE(km) | VA |
|---|---|---|---|---|
| 1 | photo | 2452.24 | 93.49 | 0.64 |
| | video | 2744.00 | 185.10 | 0.62 |
| 2 | photo | 5243.38 | 5738.70 | 0.50 |
| | video | 5719.46 | 6374.11 | 0.50 |
| 3 | photo | **2451.53** | 94.23 | 0.64 |
| | video | 2744.00 | 185.09 | 0.63 |
| 4 | photo | 2457.29 | **82.23** | **0.64** |
| | video | **2703.20** | **114.48** | **0.63** |

Table 2: Estimation & Verification results for runs

## 5. CONCLUSION AND FUTURE WORK

In this paper, we study the problem of predicting coordinates for multimedia objects. We adopt an approach which identifies the tags which are frequently used by users at each location. This, in turn helps us predict the cell and thereafter the coordinates for each object. Our analysis of wrong prediction reveals that true cells are often present is in top-k and are close to the predicted cell. This seems to be an area for improvement, where one needs to disambiguate between the neighbouring cells, maybe considering cells of varying sizes or forming clusters based on the closeness of training instances.

## 6. REFERENCES


[1] J. Choi, C. Hauff, O. V. Laere, and B. Thomee. The placing task at mediaeval 2016. *MediaEval 2016 Workshop*, Oct. 20-21 2016.

[2] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise.

[3] C. F. Karney. Algorithms for geodesics. *Journal of Geodesy*, 87(1):43–55, 2013.

[4] G. Kordopatis-Zilos, S. Papadopoulos, and Y. Kompatsiaris. Geotagging social media content with a refined language modelling approach. In *Pacific-Asia Workshop on Intelligence and Security Informatics*, pages 21–40. Springer, 2015.

[5] J. Langford, L. Li, and A. Strehl. Vowpal wabbit. *URL https://github. com/JohnLangford/vowpal wabbit/wiki*, 2011.

[6] A. Makazhanov, D. Rafiei, and M. Waqar. Predicting political preference of twitter users. *Social Network Analysis and Mining*, 4(1):1–15, 2014.

[7] P. Serdyukov, V. Murdock, and R. Van Zwol. Placing flickr photos on a map. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 484–491. ACM, 2009.

[8] H. Tamura, S. Mori, and T. Yamawaki. Textural features corresponding to visual perception. *IEEE Transactions on Systems, Man, and Cybernetics*, 8(6):460–473, 1978.

[9] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

[10] M. Wick and C. Boutreux. Geonames. *GeoNames Geographical Database*, 2011.